academy

Analyse exploratoire des données

Décision d'expansion commerciale à l'International

Présenté par Thiery KAPPE

Le contexte

En tant que Star Up de la EdTech, academy propose des formations en ligne pour le lycée et l'université.
Elle envisage se développer à l'international et souhaite s'appuyer sur les données de la banque mondiale sur l'éducation afin d'orienter sa décision d'expansion sur des pays à fort potentiel de développement commercial.

Nature et source des données collectées

Les données proviennent de l'organisme EdStats de la banque mondiale. Ces données regroupent des indicateurs internationaux sur l'éducation tels que :

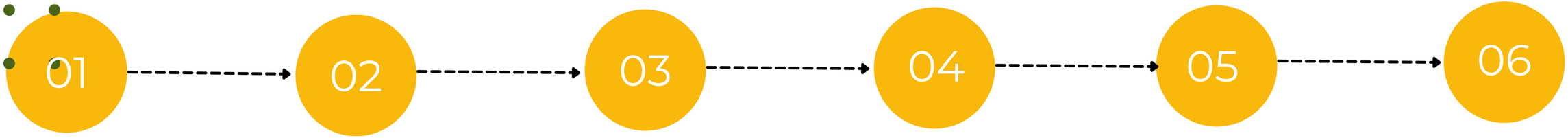
- L'accès à l'éducation,
- Le niveau d'alphabétisation,
- L'obtention des diplômes,
- Les dépenses publiques liées à l'éducation,

Les Objectifs de notre analyse

Nos travaux ont consisté à :

- Description et compréhension des blocs de variables disponibles,
- Validation de la qualité des données,
- Identification et choix des indicateurs pertinents pour notre problématique d'expansion,
- Détermination des ordres de grandeur des indicateurs sélectionnés (présentation et visualisation),
- Identification des pays à fort potentiel de développement.

Vue globale des étapes clés



Etape 1

Importation des données,
Description des blocs de variables,
Interdépendance,
Choix des blocs,
Description des variables,
Choix des indicateurs.

Etape 2

Extraction des données,
Analyse et observation globale (dimension data frame, type de variable,...)

Etape 3

Analyse exploratoire des données (doublons, données anciennes, variables sans observations),
Analyse des données manquantes,
Analyse du taux de remplissage par pays et par indicateur.

Etape 4

Analyse univariée (mesures de tendance centrale),
Visualisation graphique,
Visualisation des valeurs extrêmes,
Visualisation des top 10 des pays par indicateur.

Etape 5

Analyse bivariée,
Recherche de corrélation,
Visualisation graphique.

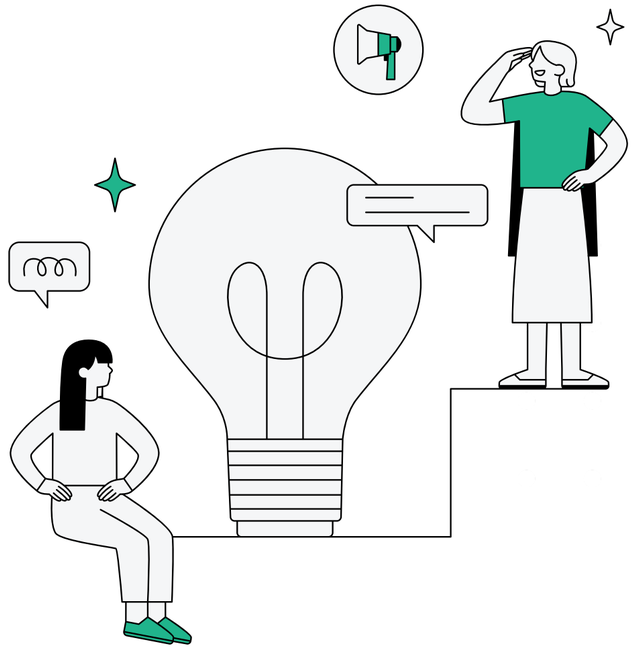
Etape 6

Mesure d'attractivité (normalisation, hiérarchie des indicateurs, poids des indicateurs),
Choix et visualisation Flop10 et Top10 des pays pour l'ensemble des indicateurs.

01

Importation des données. Description des blocs de variables Interdépendance entre les blocs

data frame	Description
country	*Il sert de base en fournissant des informations de base sur les pays et les territoires. Chaque pays est identifié par un code unique.*
countrySeries	*Il relie les pays aux séries de données spécifiques en utilisant ces codes de pays. Cela permet de savoir quelles séries de données sont associées à chaque pays.*
data	*Il associe les pays et les années aux valeurs numériques des indicateurs éducatifs. Ces indicateurs sont spécifiés par des codes de séries de données, créant ainsi un lien entre les pays, les années et les données.*
footNote	*Le data frame footNote fournit des notes explicatives qui accompagnent les données éducatives. Ces notes sont liées à des séries de données spécifiques, permettant ainsi de comprendre les détails des données.*
series	*Le data frame series contient des informations détaillées sur chaque série de données éducatives, y compris des descriptions, des thématiques, des codes uniques et des unités de mesure. Cela aide à interpréter les séries de données utilisées dans les autres blocs de variables.*



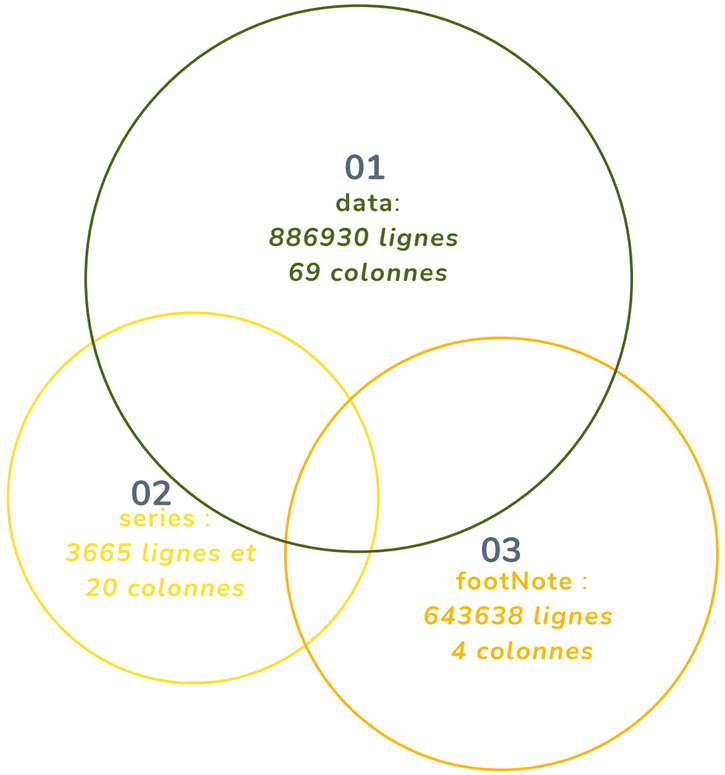
01 *Choix des blocs pertinents. Description des variables.*

Compte tenu des blocs des données et des informations qui y sont présentées, il n'est pas nécessaires d'utiliser tous les blocs de données disponibles. Nous choisissons d'utiliser les blocs data, series et footNote pour notre étude.

- ***data*** *pour la présence de l'essentiel des données numériques quantitatives sur les différents indicateurs pour l'ensemble des pays.*
- ***series*** *pour les données qualitatives sur les l'ensemble des thématiques traités.*
- ***footNote*** *pour les notes explicatives, les sources et les limitations.*

Les informations plus détaillées sur les différentes variables contenues dans ces trois data frame sont fournies dans le Notebook qui soutient cette présentation.

Concernant les dimensions voir ci-contre.



01 *Choix des indicateurs.*

Compte tenu de l'activité de academy (la formation), du mode de diffusion (en ligne), de la cible de son marché (lycée et université), nous avons réalisé des filtres à partir du data frame ***series*** *afin d'extraire les indicateurs traitants des thématiques :*

- *De l'accessibilité technique au cours* ***"Infrastructure communications"***
- *Le financement public lié à l'éducation :* ***"Expenditures"***
- *La capacité à suivre les formations dispensées, l'alphabétisation de la population :* ***"Literacy"***
- *Couverture et efficacité du système éducatif du pays via le taux de scolarisation :****"Enrolment ratio"***

Ces différents critères nous permettent d'identifier sept premiers indicateurs pertinents

Series Code	Topic	Indicator Name
IT.CMP.PCMP.P2	Infrastructure: Communications	Personal computers (per 100 people)
IT.NET.USER.P2	Infrastructure: Communications	Internet users (per 100 people)
SE.ADT.LITR.ZS	Literacy	Adult literacy rate, population 15+ years, both sexes (%)
SE.SEC.ENRR	Secondary	Gross enrolment ratio, secondary, both sexes (%)
SE.TER.ENRR	Tertiary	Gross enrolment ratio, tertiary, both sexes (%)
SE.XPD.SECO.PC.ZS	Expenditures	Government expenditure per secondary student as % of GDP per capita (%)
SE.XPD.TERT.PC.ZS	Expenditures	Government expenditure per tertiary student as % of GDP per capita (%)

02

Extraction des données des indicateurs choisis
Analyse et observation globale

*Après identification des critères, nous nous servons de ces derniers pour faire un mask que nous appliquons au data frame **data** afin de récupérer l'ensemble des données répondant aux critères identifiés. Ces données sont stockées dans un nouveau data frame "**dataEtude**"*

Aucune remarque particulière sur le type des variables :

- *Objet pour les variables composées de chaînes de caractères (Country Name, Country Code, Indicator Name, Indicator Code)*
- *float64 pour les autres variables regroupant essentiellement les observations en pourcentages des indicateurs par pays de 1970 à 2100.*

dataEtude :
1694 lignes
69 colonnes

03 *Analyse exploratoire des données (données anciennes, variables sans observation, recherche de doublons).*

Identification et suppression des données très anciennes non pertinentes pour notre étude:

Les années de 1970 à 2009 sont supprimées car très anciennes pour apporter des informations sérieuses pour la prise d'une décision d'expansion.

Identification et suppression des variables sans aucune observation :

Les années de 2017 à 2100 bien que présentes ne contiennent aucune observation, elles sont également supprimées.

Recherche des doublons :

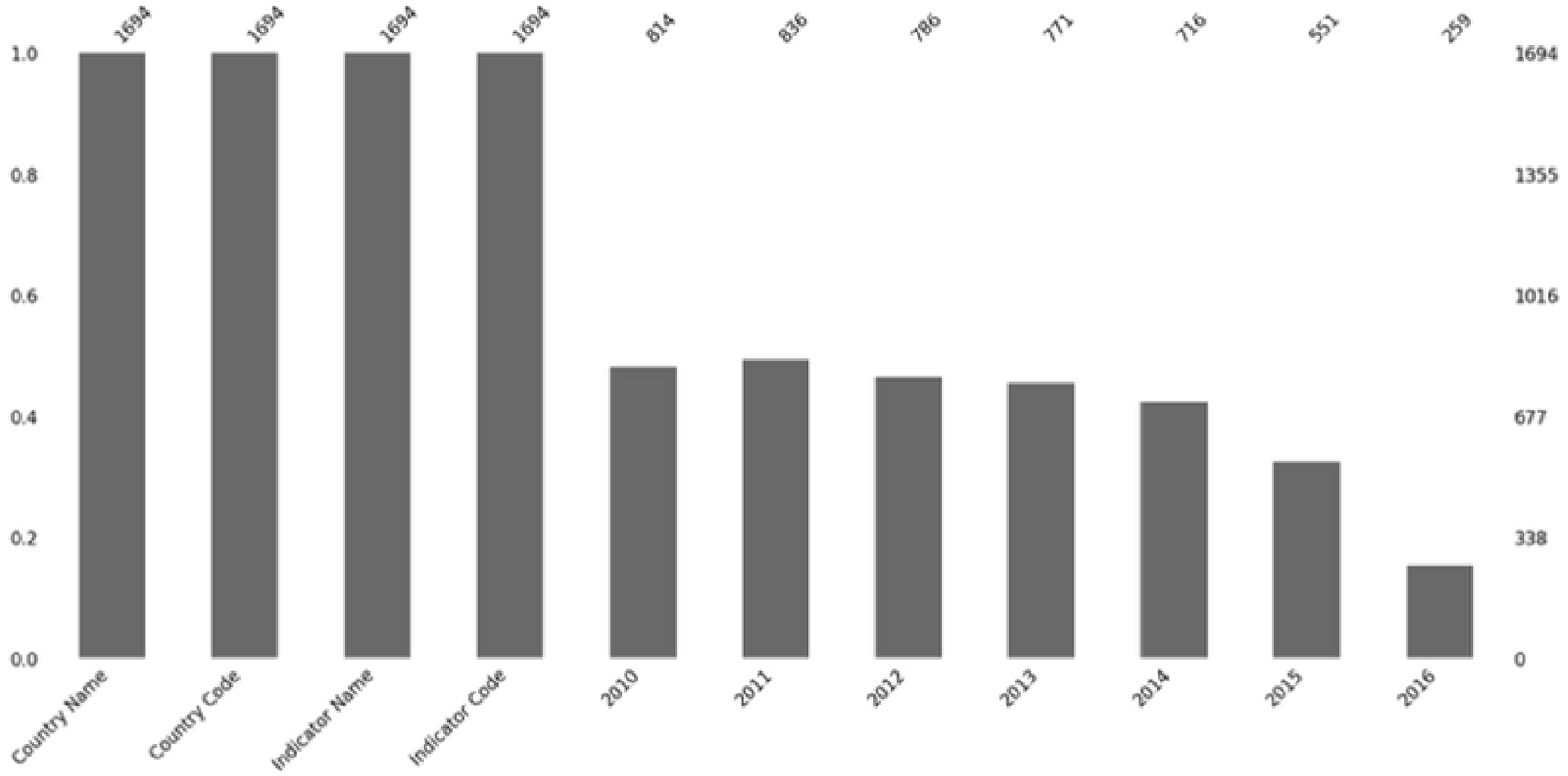
La recherche des doublons ne présente aucune anomalie.
Aucun pays n'a plusieurs fois le même indicateur.

Après ces différentes opérations, nous stockons les données dans un nouveau data frame "df"

df :
1694 lignes
11 colonnes

03 *Analyse exploratoire des données (données manquantes)*

- ***Visualisation globale des valeurs manquantes via Missingno***



df :
1694 lignes
11 colonnes

03 *Analyse exploratoire des données (taux de remplissage par pays, traitement des valeurs manquantes, taux de remplissage par indicateur)*

- ***Analyse du taux de remplissage par pays et suppression des pays dont le taux de remplissage est inférieur à 45%.***
  ==> *l'indicateur "Personal Computer" sort ainsi de la liste des indicateurs*

- ***Traitement des valeurs manquantes***

Les indicateurs identifiés comme pertinents pour la problématique d'expansion de academy, traite des thématiques liés à l'IT concernant l'accès aux formations et d'autres thèmes comme la scolarisation et l'alphabétisation. Ce sont des sujets en constante évolution au fil des années. Nous prenons donc comme postulat de ***remplacer les valeurs manquantes par la dernière valeur connue de ces indicateurs sur les années d'étude (2010 à 2016).***

- ***Analyse du taux de remplissage par indicateur après pivot***

Afin de présenter les indicateurs en colonnes pour faciliter nos traitements, nous avons effectué un pivot.
Après analyse du taux de remplissage par indicateur, nous supprimons ceux ayant un taux inférieur à 70%.

04 *Analyse univariée (tendances centrales)*

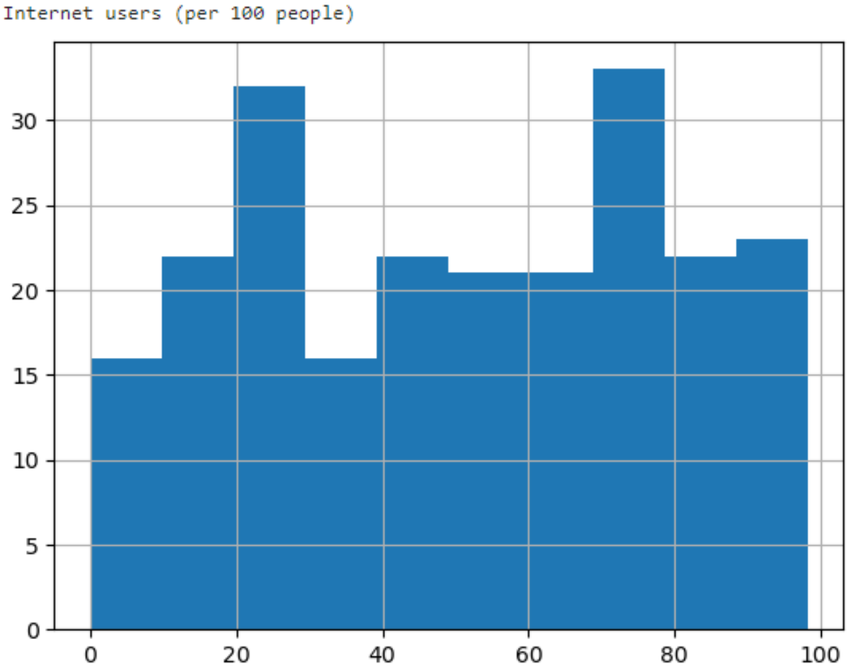
- ***Observation des tendances centrales***

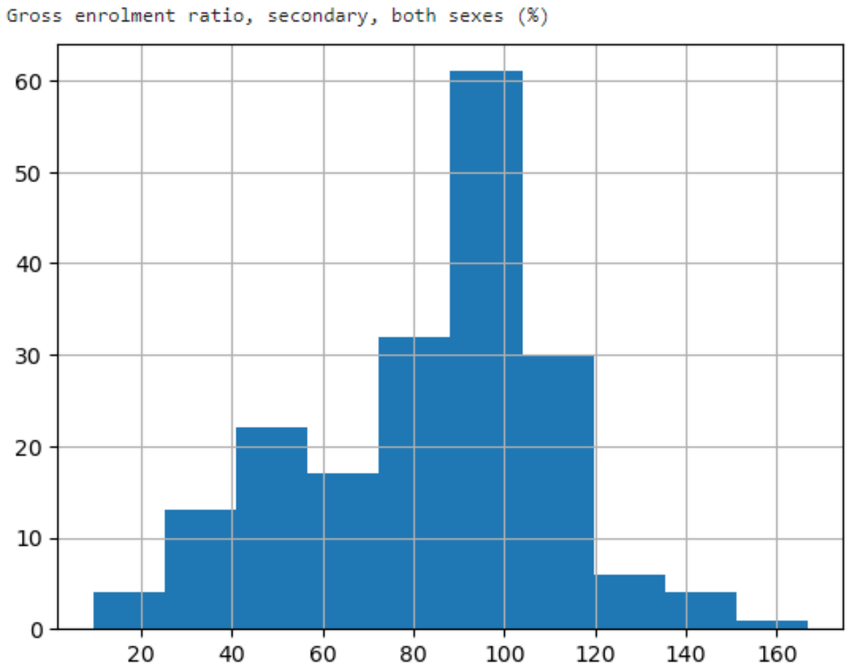
Après les traitements sur les taux de remplissage, nous avons une liste de trois indicateurs au final. Voici les tendances centrales:

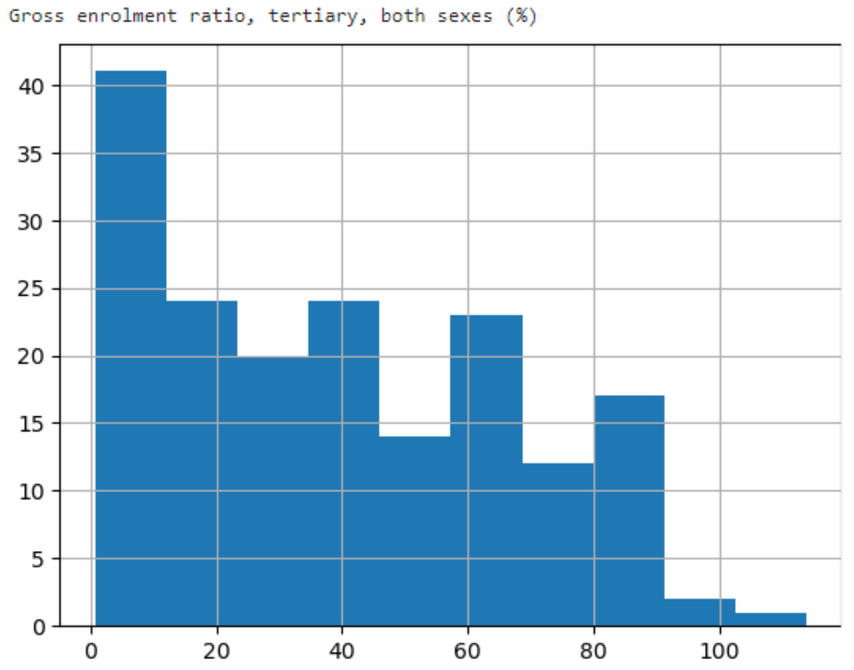
Indicator Code	Internet users (per 100 people)	Gross enrolment ratio, secondary, both sexes (%)	Gross enrolment ratio, tertiary, both sexes (%)
count	228.000000	190.000000	178.000000
mean	50.875758	84.106253	39.598241
std	28.032564	28.151614	27.611285
min	0.000000	9.517560	0.797730
25%	25.471509	65.727049	13.462642
50%	53.213089	90.366844	36.351574
75%	75.598968	102.229568	63.288498
max	98.240016	166.808472	113.871788

04 *Analyse univariée (visualisation graphique)*

- ***Histogramme de la dernière valeur connue pour chaque indicateur sur la période de 2010 à 2016.***

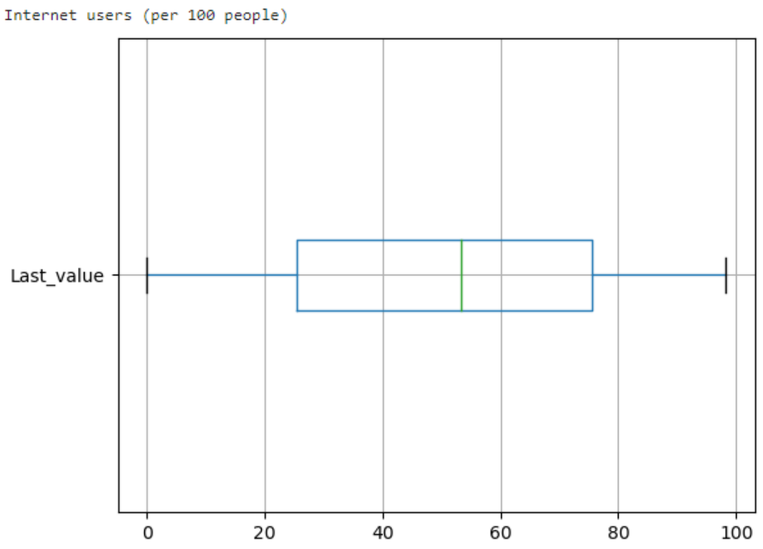


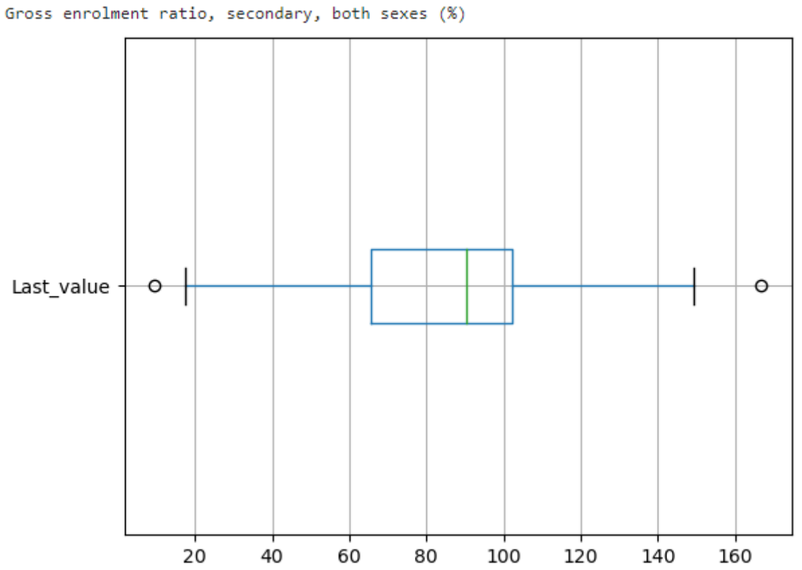


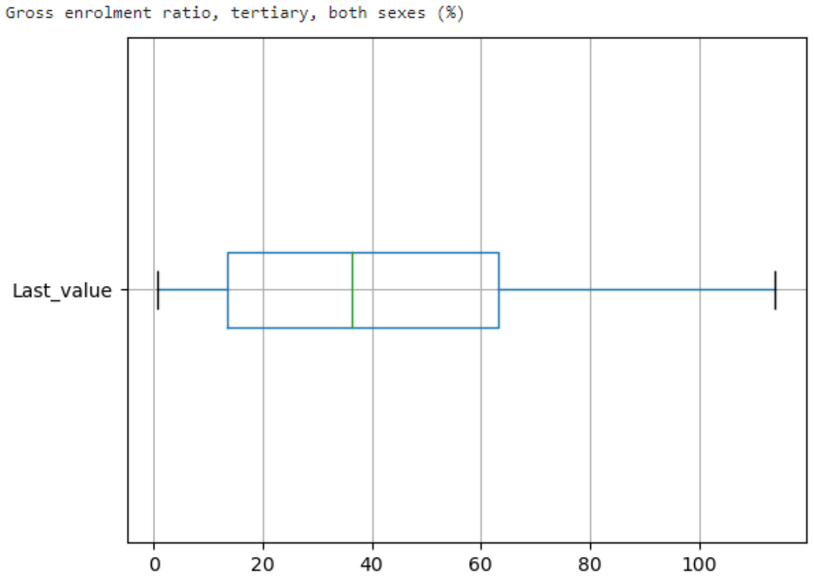


04 *Analyse univariée (visualisation des valeurs extrêmes)*

- ***Boxplot des valeurs extrêmes pour chaque indicateur.***

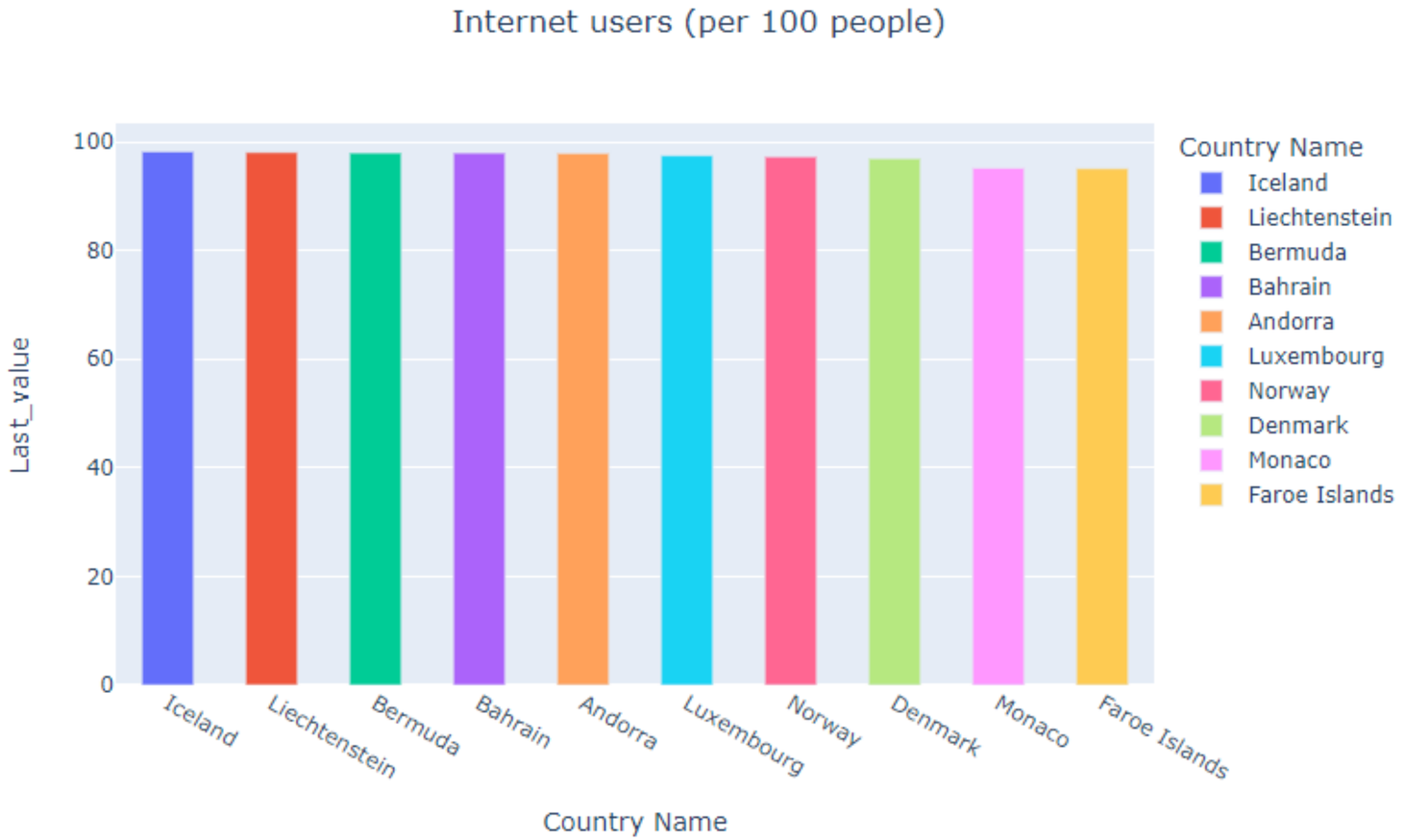


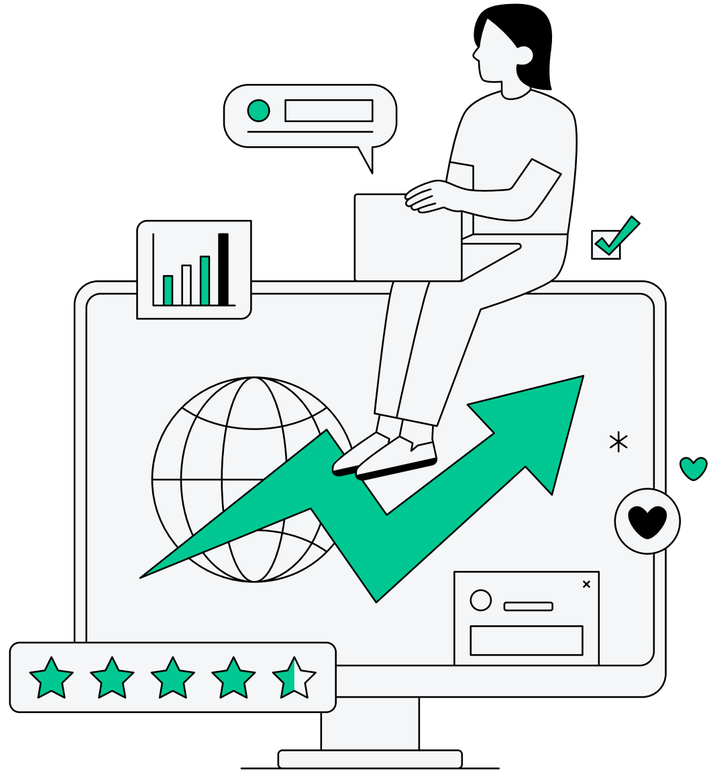




04 *Analyse univariée (visualisation des Top 10 des pays pour chaque indicateur)*

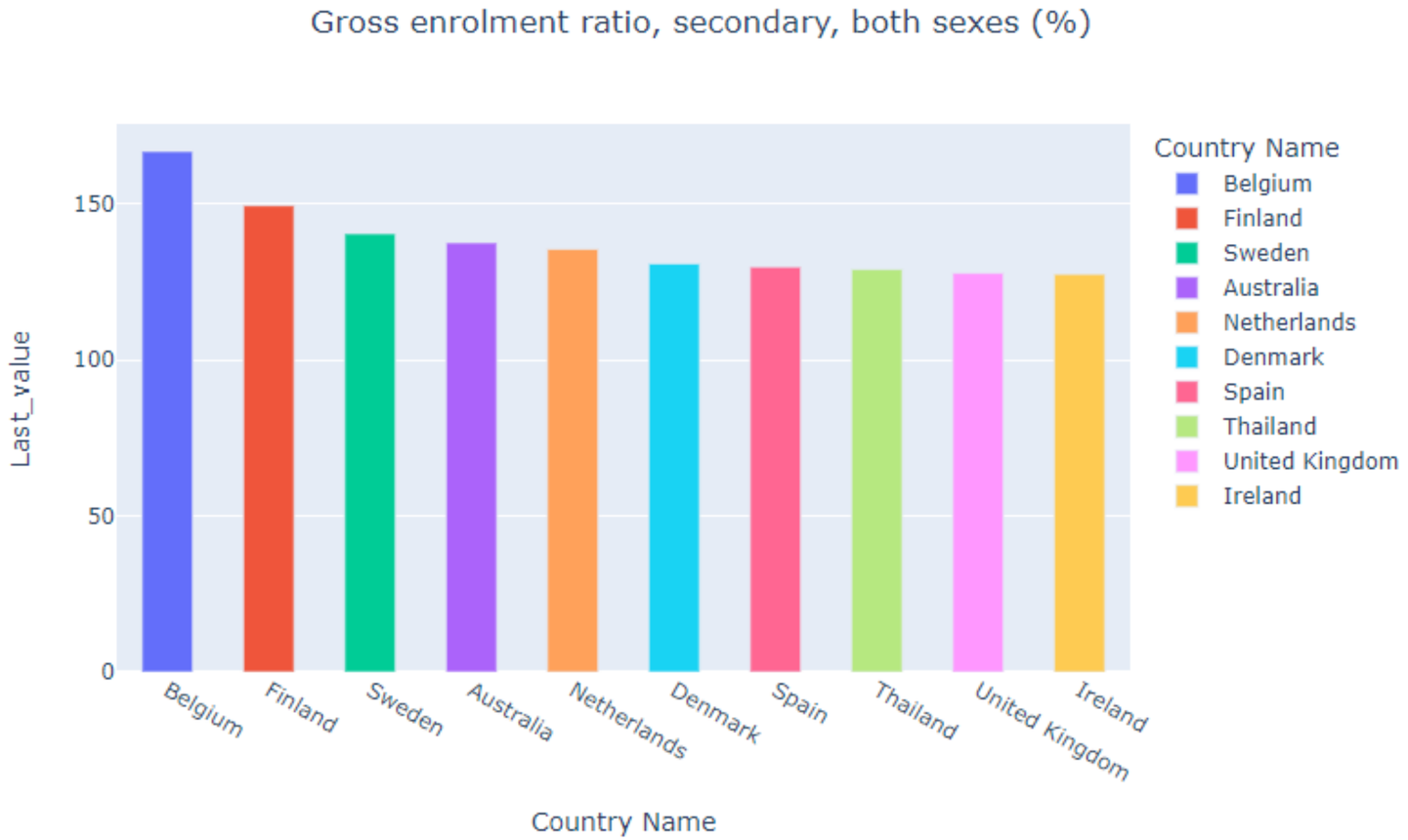
- ***Top 10 des pays pour chaque indicateur***

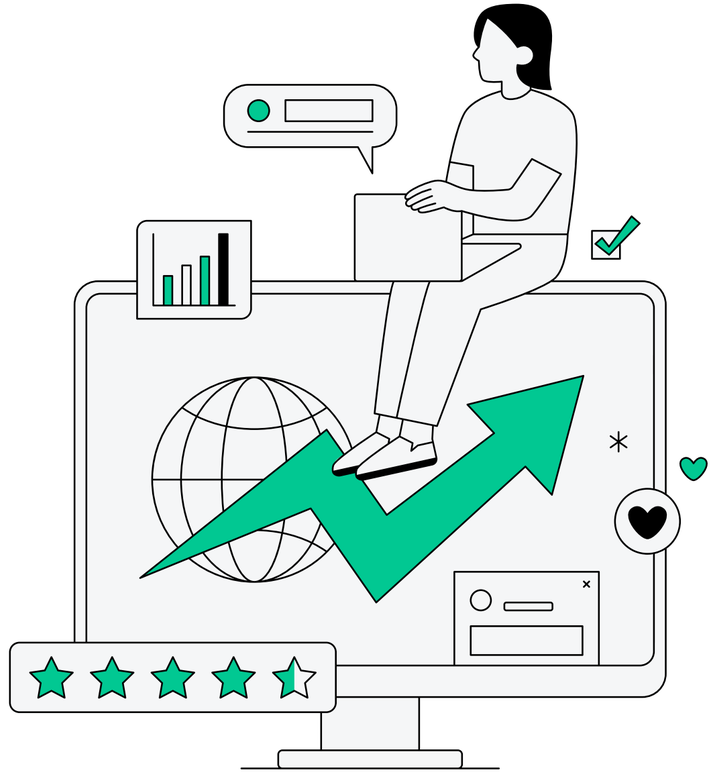




04 *Analyse univariée (visualisation des Top 10 des pays pour chaque indicateur)*

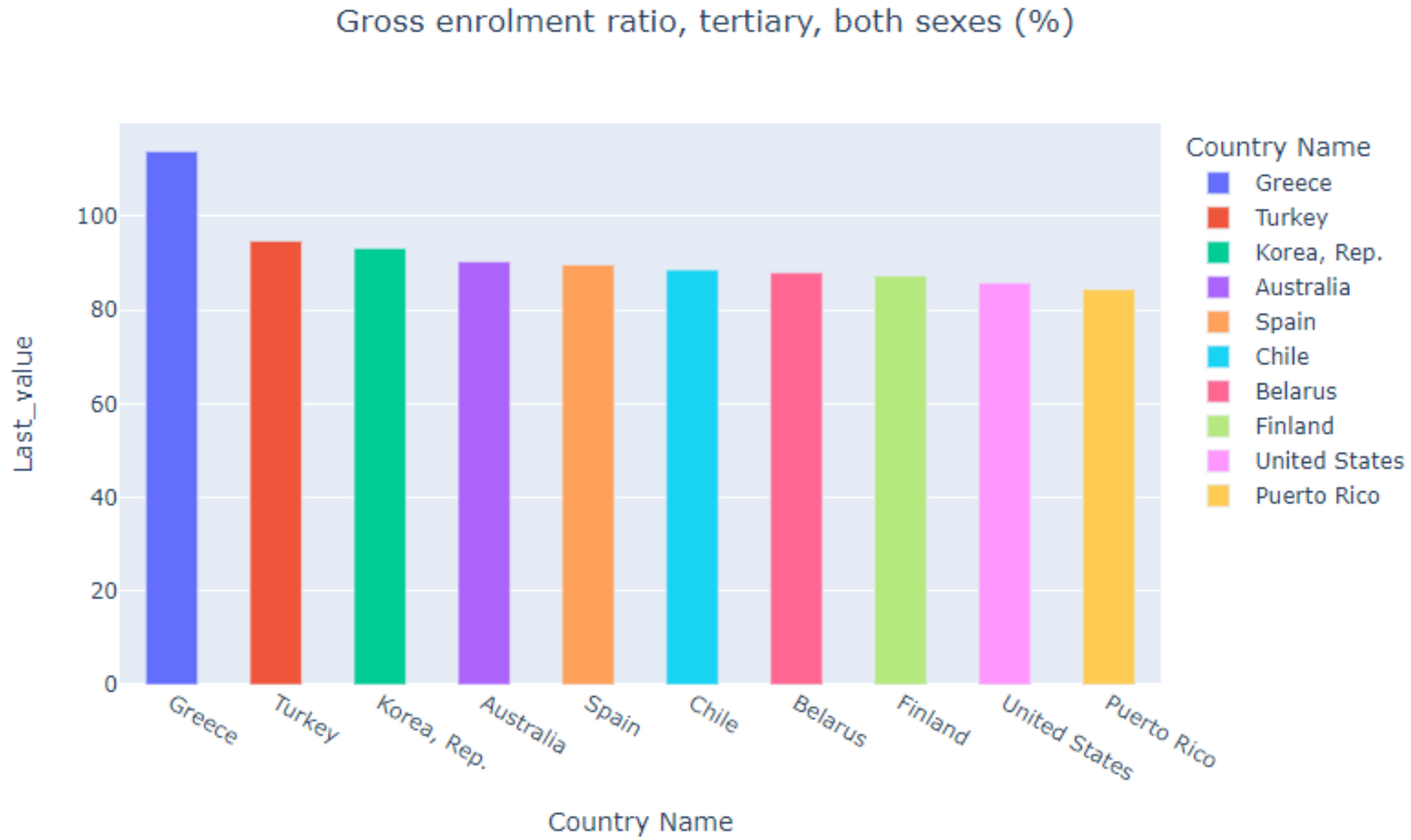
- ***Top 10 des pays pour chaque indicateur***

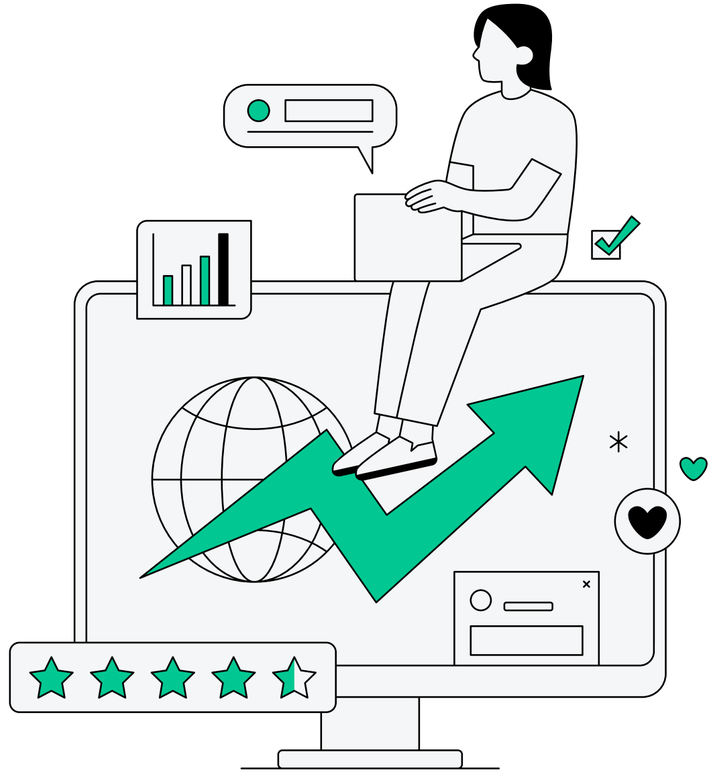




04 *Analyse univariée (visualisation des Top 10 des pays pour chaque indicateur)*

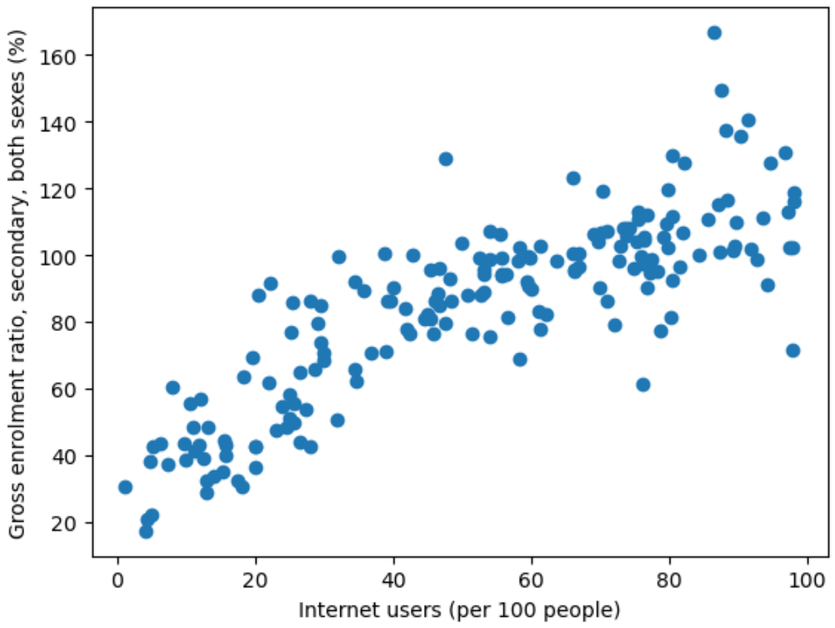
- ***Top 10 des pays pour chaque indicateur***

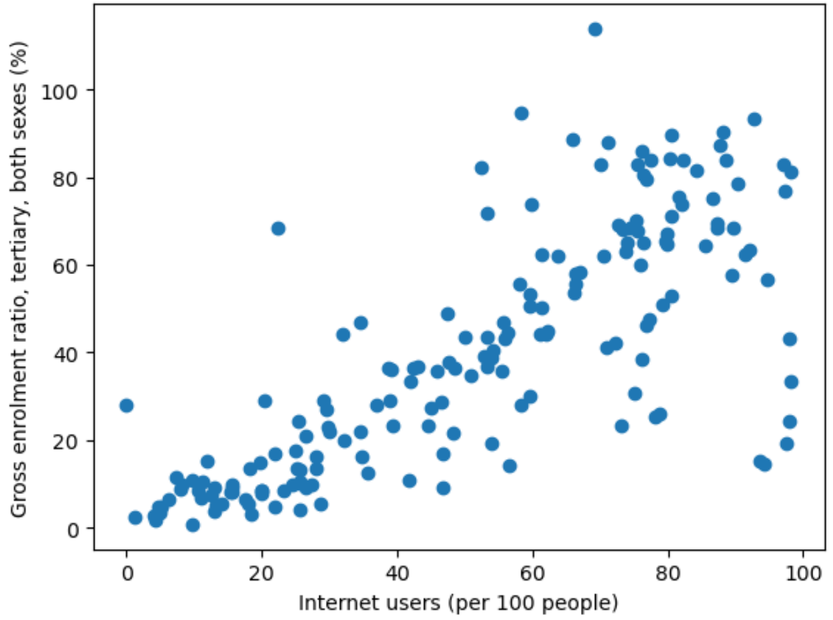


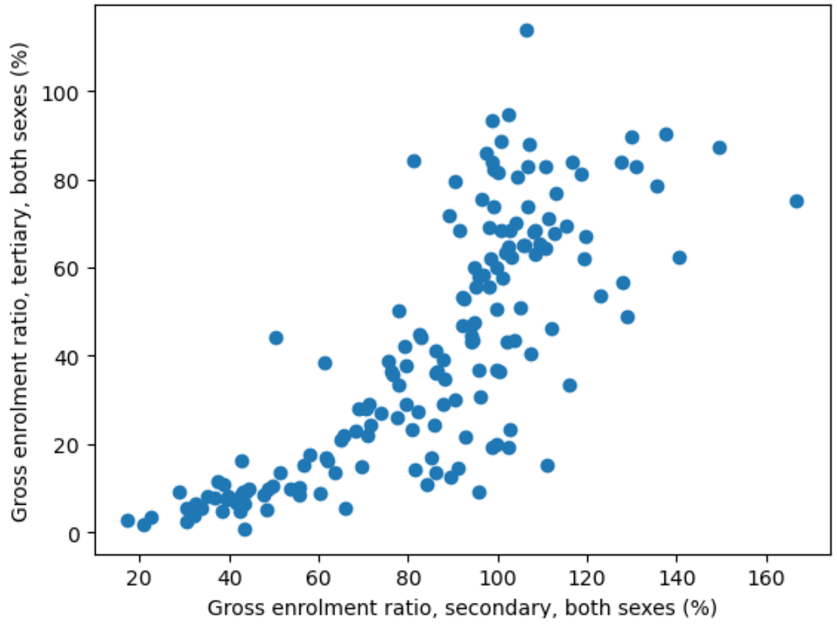


05 *Analyse bivariée (corrélation et visualisation)*

Indicator Code	Internet users (per 100 people)	Gross enrolment ratio, secondary, both sexes (%)	Gross enrolment ratio, tertiary, both sexes (%)
Indicator Code			
Internet users (per 100 people)	1.000000	0.837792	0.772925
Gross enrolment ratio, secondary, both sexes (%)	0.837792	1.000000	0.788445
Gross enrolment ratio, tertiary, both sexes (%)	0.772925	0.788445	1.000000







06 *Mesure d'attractivité (normalisation, hiérarchisation)*

- ***Normalisation des données***

Les indicateurs ne sont pas tous sur la même échelle, il est donc nécessaire de les normaliser avant de nous appuyer dessus pour réaliser des sélections.
Nous utilisons pour cela la méthode MinMaxScaler.

- ***Hiérarchisation des indicateurs***

Compte tenu de la nature de l'activité de academy et de son mode de diffusion, les indicateurs n'ont pas la même importance. Voici la hiérarchie que nous définissons :
==> Niveau 1 : Internet users (per 100 people)
==> Niveau 2 : Gross enrolment ratio, secondary, both sexes (%)
==> Niveau 2 : Gross enrolment ratio, tertiary, both sexes (%)

06 *Mesure d'attractivité (poids relatif des indicateurs, poids des pays et choix)*

- ***Attribution des poids relatifs pour chaque niveau***

==> Niveau 1 : 5 points soit un poids relatif de 5/9

==> Niveau 2 : 2 points soit un poids relatif de 2/9

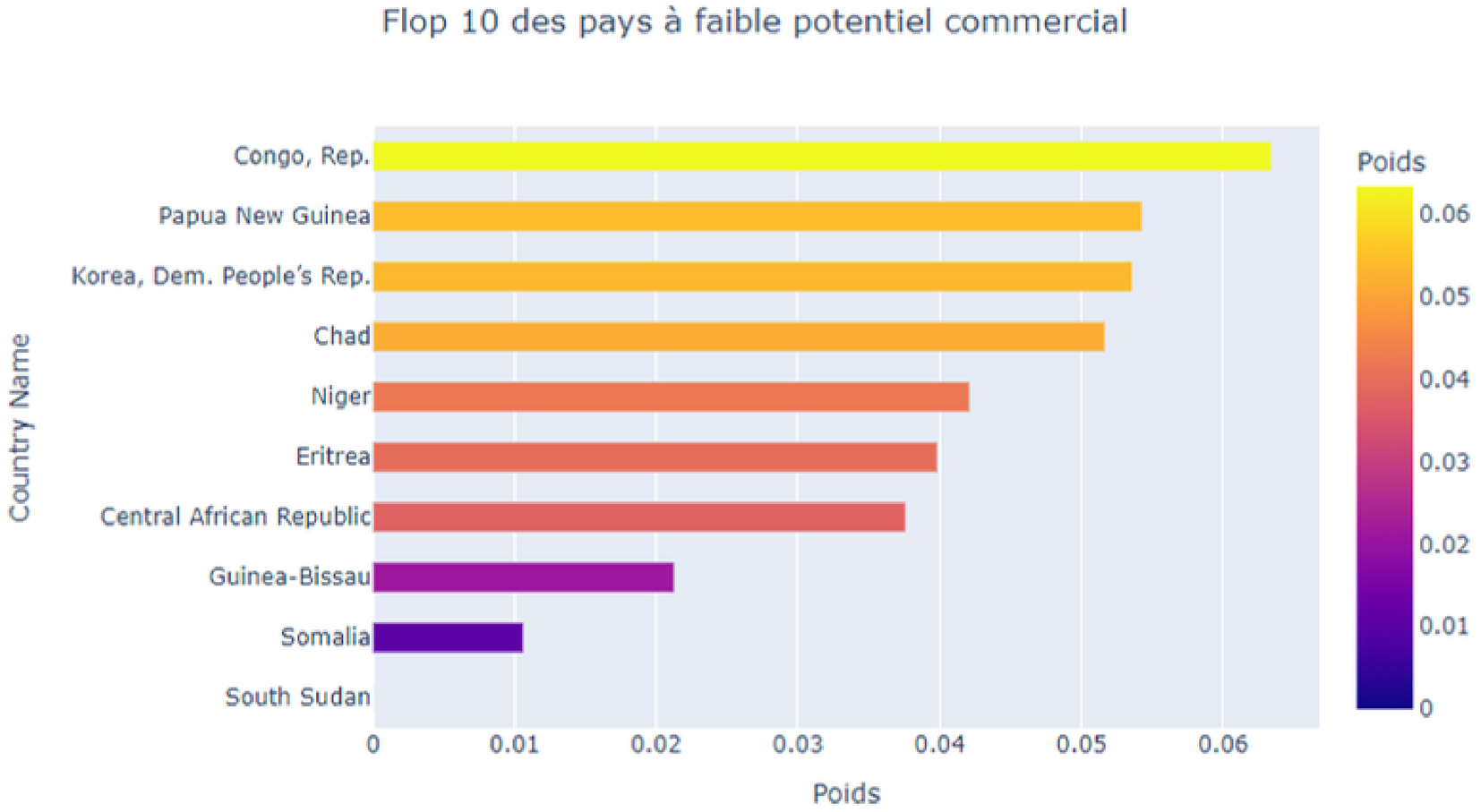
- ***Poids de indicateurs***

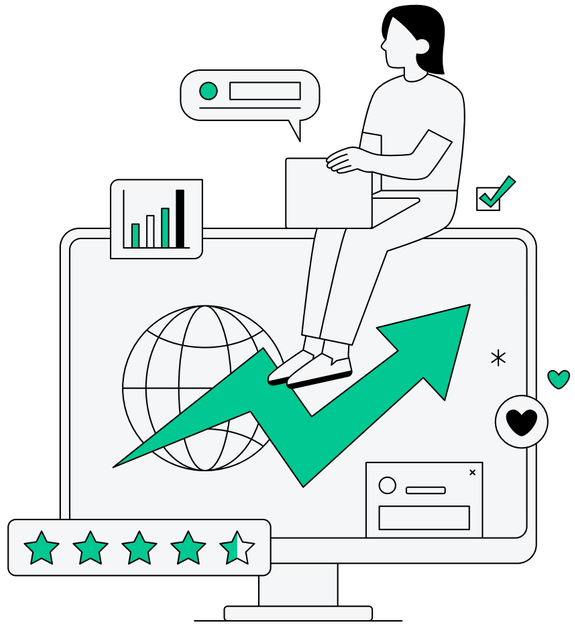
*Le poids de chaque indicateur = Last_value normalisée * Poids relatif*

Un regroupement par pays en agrégeant par la somme, nous permet de déterminer le poids de chaque pays pour l'ensemble des indicateurs.

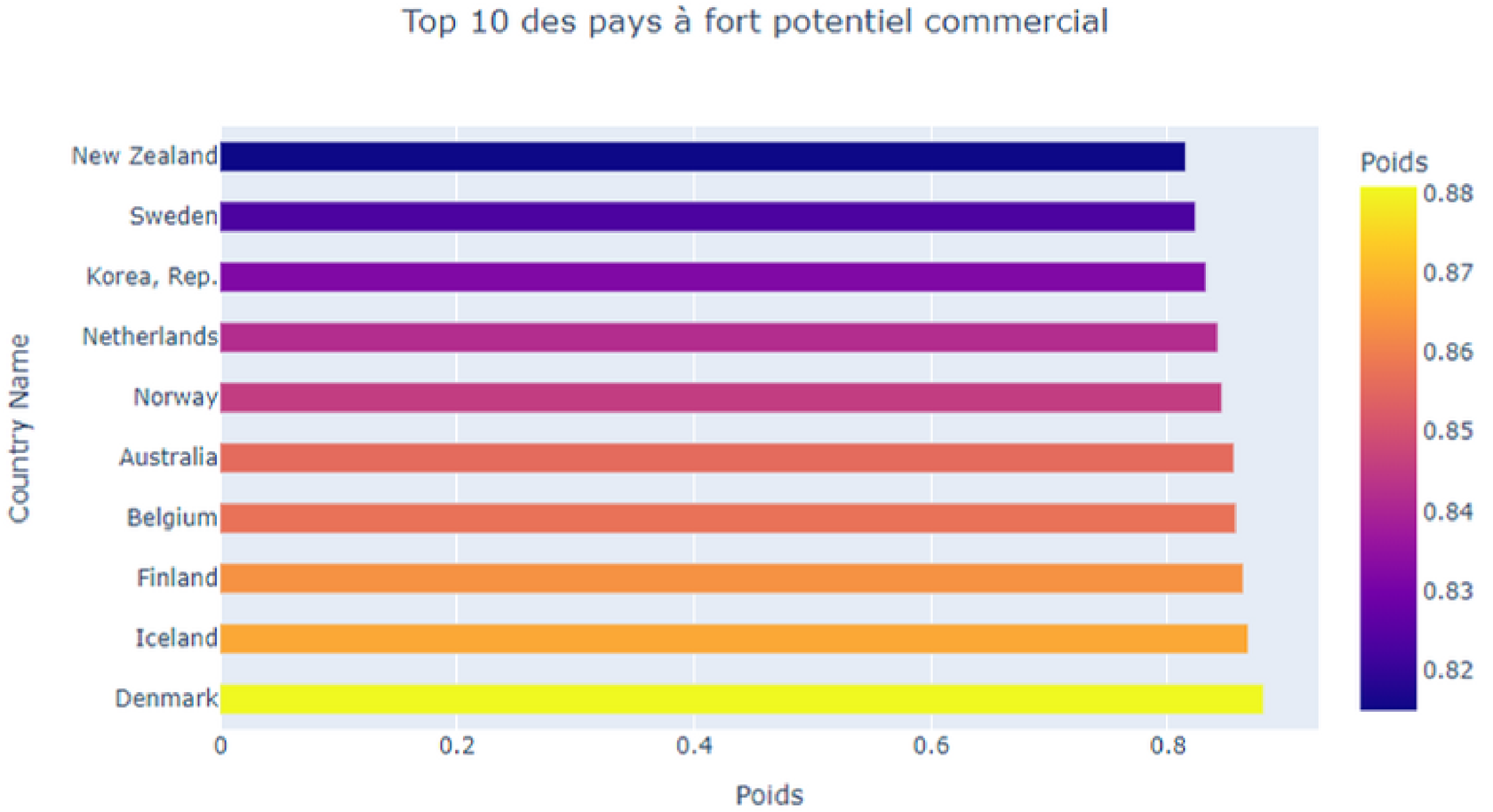
Les pays ayant le poids le plus élevé sont ceux ayant le potentiel commercial le plus élevé.

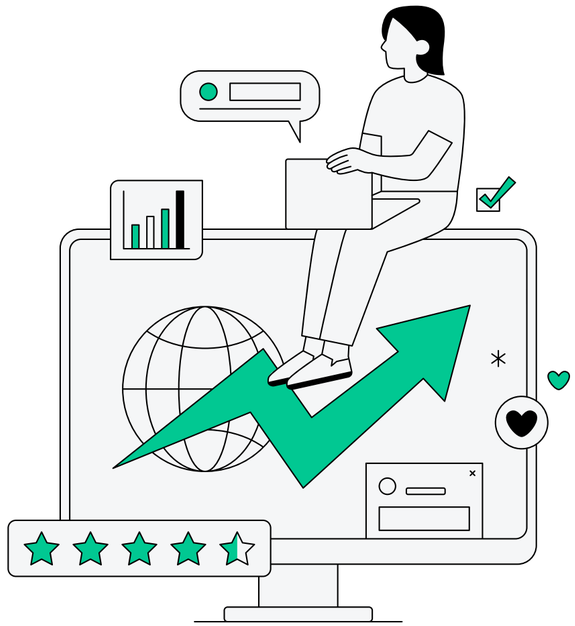
06 *Mesure d'attractivité (visualisation des Flop 10)*





06 *Mesure d’attractivité (visualisation des Top 10)*





Merci